A Practical Guide for Technical Leaders: AI Governance

Special Interest Group

ISACA
Sydney Chapter

A Practical Guide for Technical Leaders: AI Governance

Special Interest Group (SIG) Stream 02
AI Governance, Risk & Ethical Considerations

Authors (SIG Stream 02)

Varun Pant (Stream 02 Lead, Chief Editor)
Aftab Rizvi, PhD
Asaf Ahmad
Emile Ghadiminejad
George Sarandrea
Julian Petrich
Suzanne Theron

In loving remembrance of Denny Wan

Peer Reviewers (SIG Stream 01)

Wilson Chiu (Stream 01 Lead)
Krishna Bagla
Narayana Madineni
Patrick Meek

Industry Reviewers

Bharat Bajaj (ISACA Melbourne Chapter)
Kavita Chetty (ISACA Auckland Chapter)
Manpreet Singh
Maryam Shoraka

This initiative was run under the auspices of ISACA Leadership Team

Chirag Joshi (President, ISACA Sydney Chapter, 2025)
Muralee Krishnan (Coordinator, SIG 2)
Shalbin Samuel (Coordinator, SIG 1)

Table of Contents

A Practical Guide for Technical Leaders: AI Governance 1

1. Abstract 1
2. Executive summary 1
3. Introduction 2
4. Senior Management role in AI governance 3
5. Technical (and security) leader's role in AI governance 4
6. Senior Management and AI risk management 5

6.1. Framing AI risks as core business risks 5
6.2. Role of Senior Management in promoting a culture of responsible AI 6

7. AI Risks to Consider - the Baker's Dozen 8
8. Risk Management across the AI lifecycle 16

8.1. Inception & strategy 17
8.2. Design & development, or procurement, depending on the role 18
8.3. Verification & validation 20
8.4. Deployment 21
8.5. Operation & monitoring 22
8.6. Retirement 23

9. Establishing accountability and an action plan 25

9.1. The AI governance RACI matrix 25

10. Conclusion 27

Appendix: 1

A1. The First 90 Days in practice: The Leader's handbook 1

Month 1: Assess & discover (Days 1-30) 1
Month 2: Build the Foundation (Days 31-60) 1
Month 3: Implement & enforce (Days 61-90) 2

A2. Beyond the 90 days- Ongoing AI Solution Considerations 2
A3. AI procurement and securing the AI supply chain 3

A3.1 AI-Specific vendor due diligence 3

Governance, risk, and compliance considerations 4
Data provenance, handling, and privacy considerations 4
Transparency, fairness, and explainability considerations 5

Security and incident response 5
A3.2 The AI Bill of Materials (AI BOM) 5
A3.3 Contractual safeguards and liability allocation 7
A4. Deployer vs Developer considerations 7
A5. Glossary of key terms 11
CASE STUDIES 13
A6. Case study: The Air Canada chatbot involving Jake Moffatt 13
A7. Case study: The Chevy dealer's "$1 Chevrolet Tahoe" 14
A8. Case study: Female applicants downranked by Amazon recruiting tool 15
A9. Case study: Mobley v. Workday, Inc lawsuit over alleged AI hiring bias 16
A10. Case study: Clearview AI algorithm built on photos scraped from social media profiles without consent 17
A11. Case study: ChatGPT implicated in Samsung data leak of source code 18
A12. Applying STRIDE to the AI Lifecycle 20
A13. External references and further suggested reading: 21



Target Audience

It is recommended that the audience of this guide has some familiarity with Artificial Intelligence (AI) principles and concepts, as well as the concepts of business strategy and project management.



Disclaimer:

Generative AI was used to help revise this document for improved readability.

1. Abstract

A Practical Guide for Technical Leaders¤ (this guide) targets small to medium-sized enterprises (SME) seeking to realise the benefits of AI while operating within their defined risk appetite. Unlike larger corporations with extensive resources, SMEs often face challenges in implementing AI safely and effectively.

This guide identifies the top 13 AI-related risks for businesses that needs consideration before and during adoption. Without strong foundational controls, AI systems can surface sensitive content, propagate mislabelled or outdated information, and create outputs that become new risk vectors.

This further provides a comprehensive and actionable AI governance framework for technical or security leaders, such as a Chief Technology Officer (CTO) or Chief Information Security Officer (CISO), to help them implement responsible AI practices aligned with business goals noting that not all organisations will have the same title.

2. Executive summary

The widespread availability of AI has shifted the business challenge from accessing the technology to controlling it. Effective AI governance, however, is not merely about placing limits on models; it is about ensuring the integrity, accountability, and security of the data ecosystems that power them.

This guide is intended for leaders seeking to harness AI's transformative potential for a competitive advantage while maintaining safety and regulatory compliance. It embeds risk management into every stage of the AI lifecycle, from initial concept to final decommissioning to help the technical leaders navigate the complexity of use of AI, offering practical guidance to classify risks, architect security controls, manage vendors, and communicate effectively with executive leadership.

Technical and security leaders play a pivotal role as translators, articulating complex technical AI risks such as algorithmic bias or data poisoning into the clear business implications including financial loss, legal liability, and reputational damage. This translation enables the Senior Management and executive leadership to make an informed decision.

In an ever-evolving evolving regulatory landscape, effective AI governance is vital. A single AI deployment not done right could lead to significant reputational damage, regulatory penalties, and material threats to organisational business continuity.

¤The reader is expected to be familiar with standard business and cybersecurity terminology, including the role of a CTO/CISO, concepts like model drift and data drift. adversarial attacks on AI like data poisoning, model inversion attacks, and Data Loss Prevention (DLP).

Putting an AI lifecycle model into practice requires a granular, data-centric approach. A practical framework for execution must focus on data and begins with the capability to discover and classify all sensitive and business-critical data, wherever it resides. It requires the ability to automatically enforce data governance policies (e.g., permissions, sharing settings) and continuously monitor data usage to detect risks. This data-centric foundation enhances Data Loss Prevention (DLP), ensures regulatory compliance, and allows for secure integration with AI tools.

This practical, data-focused execution is reinforced by four core organisational pillars:

- **Culture of responsibility:** A strong risk-aware culture, initiated and sponsored by Senior Management, is non-negotiable. Senior Management must provide the mandate and sponsorship needed for sustainable success and champion AI literacy, approve a clear AI Acceptable Use Policy, commit to continuous employee training [Δ] and endorse consistent communication of AI risks in clear business terms.

- **Defined accountability:** This guide recommends that people be made accountable and that governance be operationalised through the clear definition and assignment of roles and responsibilities. There is a RACI (Responsible, Accountable, Consulted, Informed) matrix provided listing critical governance tasks, so they are owned explicitly, eliminating ambiguity and ensuring no critical tasks are overlooked.

- **Proactive governance:** Regulatory compliance should not be reactive. The guide advocates for anticipating and embedding compliance by design, aligning with evolving frameworks such as the EU AI Act and Australian regulatory initiatives. Compliance must be anticipated, mapped, and built into systems by design.

- **Phased action plan:** A structured 90-day plan transforms strategy into practice. It guides leaders through risk assessments, policy development, technical control implementation, and achieving executive buy-in.

By adopting this comprehensive, data-centric governance framework, as described in this guide, SME leaders can move beyond analysis paralysis and embrace AI confidently, driving innovation and growth within a measured, well-managed risk environment.

3. Introduction

For modern businesses, with the evolution of the AI landscape from a technological frontier to business imperative, with shrinking profit margins and competitors embracing the technology, it is a question of 'evolve or die'. The initial euphoria surrounding the breakout of generative AI with seemingly limitless potential has matured into a more complex reality by mid-2025.

Δ: AI training will not be the same for all. for some the focus will be on how the AI works covering topics like data science, statistics, coding (like Python), understanding model architectures (e.g. transformers, neural networks), data pre-processing, model validation, and deployment. For others the focus will be on what the AI does including topics like procurement, prompt engineering, identifying bias in outputs, data privacy implications, ethical use, and understanding compliance and governance frameworks.

Businesses are realising that AI is not a silver bullet, and that successful implementation requires careful planning and considerations across the entire AI lifecycle. The AI landscape is continuously evolving and maturing, such as the advances in Agentic AI, which brings with it both new opportunities and new risks.

The focus has fundamentally shifted from accessing AI to governing AI to realise its potential. This pivot demonstrates that the technological frontier has been settled; the core challenge is no longer accessing AI capabilities, which are now ubiquitous, but controlling them. As industry analyses confirm, workforce capability, regulatory complexity, and risk management have overtaken technical barriers as the key obstacles to effective deployment[1,2,3].

Those that fail to progress beyond surface-level adoption risk not only stalled growth but also regulatory sanctions and reputational damage. In this environment, long-term survival depends on mastering AI governance; not merely adopting the technology.

Many Businesses are now attempting to try balance the desire to reap the benefits of AI with the need to integrate it responsibly within their core operational processes that are increasingly sophisticated and interconnected. Embedding AI into the organisational DNA requires a disciplined, risk-aware approach.

To navigate this landscape, enterprises must adopt a structured governance framework, as described in this guide, to transform AI risk management from a compliance burden into a sustainable competitive advantage. Achieving this requires decisive Senior Management leadership, empowered technical expertise, and coordinated cross-functional participation, aligned through the AI Governance RACI Matrix in section The AI governance RACI matrix of this guide.

4. Senior Management role in AI governance

Senior Management[¥] (for example, the CEO and other C-suite executives) have a crucial role in ensuring AI is used responsibly. By approving the strategy, allocating necessary resources, and championing the initiative, Senior Management provides the mandate and sponsorship essential for success, setting the tone for responsible AI practices across the business.

To operationalise this accountability, Senior Management's role extends beyond sponsorship to include a formal, periodic management review. This structured oversight function is critical for governance and transforms their high-level responsibility into evidence-based leadership.

[¥] Senior Management is distinct from the Board of Directors. The Board governs and has oversight whilst Senior Management oversees day-to-day leadership and strategy execution. Senior Management drives the AI strategy and its real-world outcomes and acts as the decision-makers for its responsible use. They also, approve, fund, and champion the AI strategy. This top-down sponsorship is essential for establishing the non-negotiable mandate for responsible AI practices across the entire organisation.

1. The state of AI: How organizations are rewiring to capture value
2. https://www.aigroup.com.au/resourcecentre/research-economics/technology-adoption-in-australian-industry/
3. 2025 AI Deployment and Governance Survey Report

This review process must be informed by key inputs, including:

- The status of actions from previous reviews.
- Changes in the external and internal landscape, such as new technologies, competitor actions, or evolving Australian and relevant international regulatory requirements.
- Shifts in the needs and expectations of key stakeholders.
- Direct and resource the creation of a central, living AI Asset Inventory (or Model Inventory). This is the foundational tool for compliance.
- Ensure ongoing executive oversight of the AI management system's performance, including the review of audits, nonconformities, and opportunities for improvement.

By actively reviewing this information, Senior Management can proactively promote a robust data culture to address challenges like data inaccuracy, privacy, and bias. This requires a new governance structure, one that fundamentally integrates technical expertise into risk and compliance functions.

The first and most critical action is to establish a permanent, empowered AI Governance team representing all relevant decision makers and stakeholders. Additionally, since AI adoption involves significant change, the Senior Management in this review process must provide a clear, top-down mechanism for communicating objectives and directing action.

Furthermore, given the rapid evolution of AI, this agile approach to review ensures the business remains flexible and responsive. Senior Management empowers the technical and security leaders who are tasked with the day-to-day implementation and operation of the AI Governance program while still providing oversight and strategic direction.

Note that while Senior Management's role is mentioned at a high-level given their importance, they are **not** the primary audience, and this document is about the technical and security leader's role and activities to set up the AI Governance Program.

5. Technical (and security) leader's role in AI governance

The technical or security leader — such as CTO or CISO (The Leader)— in a business, are at the forefront of both harnessing the transformative power of AI to compete effectively whilst also defending the business against unacceptable levels of risks. SME must be more deliberate, more focused and more strategic in their AI Governance than the large firms, ensuring that there is a mature governance posture underpinned by a robust framework of controls, policies and oversight needed to manage AI's heightened risks (unique risks as well as existing risks amplified by the use of AI).

The role in this new landscape extends far beyond managing IT infrastructure and securing the corporate network. It demands the active governance of complex, often opaque, AI systems. The perimeter to defend has dissolved; the risks are no longer just external threats trying to get in, but are embedded within the logic, data, and behaviour of the tools the business relies on.

Responsibilities have evolved to include data provenance, model integrity, ethical oversight, and the ongoing monitoring of algorithmic behaviour.

In this context, The Leader must manifest in two distinct but interconnected roles, namely the Translator, and the Architect.

First, as a translator, The Leader must be capable of articulating complex, technical AI risks into clear, unambiguous language of business impact. Senior Management may not understand the nuances of model drift (degradation of the model's predictive capabilities), but they will act decisively when it is explained as, "The AI-driven inventory system is becoming less accurate over time, which will lead to a projected $200,000 in excess stock and wasted capital this year." They will grasp the threat of data poisoning when it is framed as, "A competitor could deliberately feed our public survey tool with false information, causing misallocation of the entire quarterly marketing budget."

The ability of The Leader to convert abstract threats like algorithmic bias into concrete liabilities such as costly discrimination lawsuits and irreversible reputational damage is essential for securing the mandate and resources required.

Second, as an architect, The Leader must be capable of designing and implementing the business's capacity for responsible innovation from the ground up. This is a hands-on, structural role that includes the following:

- Orchestrate the oversight process: Establish a lightweight but effective AI review committee or process, ensuring that new use cases are evaluated for both business value and potential harm before deployment but also ongoing with a view to decommission if no longer delivering the expected benefits.
- Embed security and ethics: Champion the integration of security and ethical checks directly into the AI development and procurement lifecycle, ensuring that governance is not an afterthought but a foundational component of every AI initiative.
- Identify and manage AI specific risks: *Please see next section.*

By embracing these roles, The Leader moves from being a technical manager to a strategic partner, embedding responsible practices into the DNA of the organisation and ensuring that the SME can innovate with both speed and safety.

6. Senior Management and AI risk management

To secure executive buy-in and appropriate resources, The Leader must articulate AI risks not as abstract technical possibilities but as concrete business threats.

The tangible business consequences of AI risks must be communicated to Senior Management, grounding the discussion in documented financial, legal, and reputational outcomes.

6.1. Framing AI risks as core business risks

Discussions on AI risks must be framed as core business risks. To help understand the nature of the risks, case studies are included in the appendix of this guide and cross referenced here.

Operational & financial damage: Decisions based on flawed or "hallucinated" AI outputs can lead to significant financial loss.

- Case study: The Air Canada chatbot involving Jake Moffatt (see appendix)
- Case study: The Chevy dealer's "$1 Chevrolet Tahoe" (see appendix)

Reputational harm: Beyond immediate financial loss, AI failures can inflict deep, lasting damage on a company's brand, customer trust, and legal standing. A single incident of a biased algorithm or a privacy breach can erode customer trust and damage the brand for years.

- Case study: Female applicants downranked by Amazon recruiting tool (see appendix)
- Case study: Mobley v. Workday, Inc lawsuit over alleged AI hiring bias (see appendix)

Data privacy & security: AI systems are data-hungry, increasing the attack surface and risk of non-compliance with regulations like the Privacy Act and GDPR.

- Case study: Clearview AI algorithm built on photos scraped from social media profiles without consent (see appendix)

Intellectual Property (IP) leakage: The most immediate and pervasive AI risk is the leakage of sensitive corporate data and IP through the unsanctioned use of public generative AI tools by well-meaning employees. Unvetted generative AI tools pose a direct threat of leaking trade secrets, source code, and strategic plans.

- Case study: ChatGPT implicated in Samsung data leak of source code (see appendix)

6.2. Role of Senior Management in promoting a culture of responsible AI

While the implementation of technical controls and governance frameworks is necessary, it will remain insufficient if disconnected from the human and cultural dimensions of the business.

The most significant AI security vulnerability often lies not in the code but in the corporate culture, set by Senior Management. The most advanced security infrastructures may be compromised not by technical flaws but by the inadvertent actions of a single, untrained employee.

A culture of responsibility must be championed from the very top of the business.

Senior Management must visibly model responsible AI use, consistently communicate the strategic importance of ethical and secure AI, and, most importantly, allocate the necessary resources for training, tools, and governance programs.

When employees see that leadership is genuinely committed to responsible AI, it ceases to be a mere compliance hurdle and becomes a shared business value.



Having written policies alone does not create culture!
The principles of acceptable use of AI must be embedded into the daily practices and mindset of the workforce through deliberate and continuous effort. Also, it is recommended to appoint a dedicated nominated person at the right level in the organisational hierarchy with the appropriate knowledge, who is accountable for the organisational use of AI.

One of the cornerstones of this cultural foundation is the AI Acceptable Use Policy (AUP) that reflects the management's intent. This document should be clear, simple, concise, devoid of technical jargon-filled legalese and rigorously enforced using effective transparent controls.

This AUP provides all employees with clarity on acceptable AI behaviour. It is designed to empower responsible innovation while ensuring all AI-driven activities operate within the business's established risk and ethical guardrails.

When drafting the AUP, the following should be included:

- a clear directive forbidding the input of any confidential company data, client information, or personal data into any public or unvetted AI tool.
- being transparent by asking permission before recording, transcribing, or using AI.
- a maintained and easily accessible list of company-vetted and approved AI services. This provides employees with safe, sanctioned alternatives for improving their productivity, channelling their desire for efficiency into secure platforms, while minimising the risk of unmanaged and shadow AI.
- a requirement for human oversight, review, and fact-checking of any AI-generated content that will be used for significant internal decisions or any external communication. This mitigates the risk of acting on AI hallucinations or biased outputs.

Effective AI literacy training moves beyond the basics of what AI is and covers the specific limitations and risks relevant to an employee's role. This includes practical education on how to spot AI hallucinations, the concept of algorithmic bias, and the security risks of sharing data. The goal is to equip employees with the critical thinking skills needed to work with AI tools responsibly and effectively.



'Soon after the first automobiles were on the road, there was the first car crash. But we didn't ban cars—we adopted speed limits, safety standards, licensing requirements, drunk-driving laws, and other rules of the road.'

– Bill Gates, cofounder of Microsoft

7. AI Risks to Consider - the Baker's Dozen

For The Leader, securing the AI transformation means looking beyond traditional cybersecurity to address new, AI-specific vulnerabilities and threats. The core risks involve the data used to train models, the models themselves, the infrastructure they run on, and the governance frameworks that control their use. The below summarises a few such risks.

1. **Sensitive data leakage:** Well meaning, but not AI literate employees, seeking productivity gains, capture or upload proprietary information—source code, financial data, strategic plans—into public generative AI systems, leading to irreversible IP loss or breaching of non-disclosure agreements. (Also, see related *Data Sovereignty and Jurisdictional risk*).
   Suggested risk treatment:
   Policy controls: *Implement and enforce a clear, unambiguous policy that explicitly forbids pasting any proprietary, confidential, or personal information into public generative AI tools. Also, manage risks related to 3rd Party/4th Party suppliers.*

   Technical controls: *Deploy private, sandboxed AI environments for any work involving sensitive data. Use Data Loss Prevention (DLP) techniques (for e.g. software and network-level blocks) to prevent data from being sent to unauthorised public AI websites, guardrails in the LLM to prevent prompt injection to disclosure sensitive data.*

   Culture and training: *Educate employees on the specific risks of IP loss. Crucially, provide employees with approved, secure internal AI tools that meet their productivity needs, thereby removing the incentive to use risky public alternatives.*

2. **Data privacy violations:** Personal information that is captured or uploaded (see risk 1) or used for model training or inference without explicit consent or a valid legal basis. Also, inadvertent memorisation and leakage of sensitive information from the training data, such as personal details or proprietary company information. Both can result in penalties under regulations like the European Union's GDPR and the Australian Commonwealth Privacy Act 1988.

   Suggested risk treatment (in addition to Risk 1):
   Privacy by Design: *Embed privacy considerations into the very beginning of any AI project. Conduct a formal Privacy Impact Assessment (PIA) as required under Australian law for any system that will handle personal information, subject to threshold analysis.*

   Legal basis and consent: *Work directly with legal and compliance teams to ensure there is a valid legal basis for processing personal data and that explicit, informed consent is obtained from individuals where required.*

   Data minimisation and anonymisation: *Adhere strictly to the principle of data minimisation—only use the data absolutely necessary for the task. Where possible, use anonymised or pseudonymised data for training to reduce privacy risks.*

3. **Algorithmic bias & discrimination:** AI systems, trained on biased data, used in, for example HR recruitment, performance management, lending, or customer service may lead to harmful discriminatory outcomes, costly lawsuits, and severe brand damage.
   Suggested risk treatment:
   Data governance and auditing*: Before training, audit datasets for biases across demographic groups (e.g., age, gender, ethnicity). Use techniques like re-sampling or data augmentation to create a more balanced and representative dataset. Establish data ownership, with proper accountability and responsibilities in place.*

   Fairness testing: *During development and before deployment, rigorously test the model's performance and outcomes across different subgroups using established fairness metrics.*

   Human-in-the-Loop (HITL): *For high-stakes decisions like recruitment, lending, or performance management, mandate that the AI system serves as a recommendation tool only. A trained human must have the final, accountable decision.*

   Transparency and explainability: *Use explainable AI (XAI) techniques to understand and document why a model made a particular decision, ensuring the logic is fair and defensible.*

   Consider avoiding systems that do not provide transparency for certain use cases, especially the ones that may cause significant harm, discrimination, or otherwise adverse impacts on individuals.

4. **Misinformation, disinformation and hallucinations:** Incorrect use of a model e.g. for a purpose not designed for, limited context in prompt or uploading data that cannot be interrogated can generate confident but incorrect AI outputs resulting in loss of trust, incorrect decisions, strategic missteps, financial losses, and reputational damages. AI models are known to hallucinate or generate outputs that are confident, plausible, but factually incorrect or entirely fabricated.

   Suggested risk treatment:

   Mandatory human verification: *Institute a non-negotiable policy that no significant business decision can be based on unverified generative AI output. Treat all outputs as a "first draft" that requires expert human fact-checking.*

   Retrieval-Augmented Generation (RAG): *Improve generative AI models to retrieve and incorporate new information and supplement information from the training data in a controlled manner, by accessing internal knowledge base of verified company data or generate responses based on authoritative sources. This forces the model to generate responses based on trusted source documents rather than its generalist, and potentially flawed, training data.*

Prompt engineering and sourcing: *Train users to prompt models in a way that encourages fact-based answers and to demand sources for any claims made. If the AI cannot cite a verifiable source, its output should be treated with extreme scepticism.*

Watermarking or disclaimers: *Get models that allow watermarking content or provide disclaimers that make it apparent that the content is AI Generated and may need to be verified.*

5. **Model degradation and drift:** Risk of model drift caused either when the input data changes over time, potentially reducing model accuracy, or a change in the relationship between the input variables (the data it uses to make predictions) and the target variable (what it's trying to predict). This results in outdated, misaligned model predictions with real world data over time, and without accountability to monitor and align model performance, it can result in inaccurate output, poor business decisions, reduced reliability, and loss of user trust.

   Suggested risk treatment:

   Continuous monitoring and alerting: *Implement automated statistical tests to continuously monitor key model performance metrics (e.g. accuracy, prediction drift) and statistical properties of the input data. Set up alerts to notify the relevant teams when performance drops below a predefined threshold.*

   Retraining cycle: *Establish a formal policy for periodically retraining the model with fresh, up-to-date data. The frequency should be based on the volatility of the operating environment.*

   Model versioning and governance: *Maintain a rigorous version control system for models. This ensures that if a newly deployed model underperforms, there is a clear and tested process to roll back to a previous, stable version with minimal disruption.*

6. **Shadow AI:** The unvetted and unmonitored use of AI tools by employees or departments creates a massive governance and security blind spot, exposing the business to all other risks on this list without any visibility or control.

   Suggested risk treatment:

   Discovery and monitoring: *Use Cloud Security Posture Management tools or other Cloud Access Security Broker (CASB)/ SaaS Management tools to discover all third-party AI applications used within the business, including those not approved by IT.*

   Establish an "approved list": *Rather than just blocking tools, create and promote a catalogue of vetted, approved, and secure AI applications that meet the business's security and compliance standards.*

Centralised governance and education: *Funnel all AI tool procurement through a central review process. Educate employees on the risks of using unvetted tools and make it easy for them to request a review of a new tool.*

7. **Copyright infringement and unauthorised use of intellectual property (IP):** Capturing or uploading copyrighted or IP owned information into an AI tool or using a generative AI tool that was trained on copyrighted material without license can result in the AI producing outputs that infringe on existing copyrights, exposing the business to legal challenges and regulatory fines.

   Suggested risk treatment:

   Vendor due diligence and indemnification: *Scrutinise potential AI vendors on their training data. Prioritise vendors that offer IP indemnification, which provides a level of legal protection if their tool generates infringing content.*

   Use of "clean" models: *For generating content where IP ownership is critical, use AI models that are exclusively trained on licensed or public domain content or fine-tune a model on the business's own proprietary data. Combine with blocking web-access to bots, and limit the search to internal, vetted sources with known data provenance.*

   Human review and modification: *Implement a policy that all AI-generated content, especially creative works or code, must be substantively reviewed and modified by a human before being published or sold. This strengthens a claim of transformative use and ownership.*

   Culture and training: *Educate employees on the specific risks of copyright infringement and unauthorised use of IP and caution against capturing or uploading copyrighted or IP owned information into AI tools.*

8. **Vendor Lock-in:** Becoming overly dependent on a single, opaque AI vendor without contractual guarantees for availability and data portability impacts business continuity and makes it difficult and costly to switch providers, reducing negotiating leverage and strategic flexibility.

   Suggested risk treatment:

   Contractual safeguards: *During negotiation, insist on contractual rights to availability of the tool but also data portability. Ensure that the contract specifies the Recovery Time and Recovery Point objectives of the AI tool, and explicit include clauses on data ownership and data portability (if the business chooses to leave the vendor and wants to move their data), and include the process, mutual responsibilities and the data extract format.*

   Platform-agnostic architecture: *Where possible, build an abstraction layer between business owned applications and the vendor's AI model. This allows the business to switch*

the underlying model provider (e.g., from OpenAI to Anthropic to Google) with minimal disruption to business owned systems.

Embrace open-source: *For key parts of the AI stack, consider using high-quality open-source models. This provides maximum control and flexibility, preventing reliance on a single commercial entity, though it requires more in-house technical expertise.*

9. **Excessive cost & complexity:** Choosing AI as a solution without defining the business problem first (where AI is considered the default solution instead just of one of the potential solutions without due considerations on fitness for purpose) often leads to selecting a solution that is overly complex and expensive for the business problem at hand, resulting in poor Return on Investment (ROI) and wasted resources. Additionally, lack of AI competency can lead to businesses selecting inappropriate, poorly performing, unsuitable or incompatible AI systems.

   Suggested risk treatment:

   Problem-first, not solution-first: *Mandate that any AI project begins with a clear business problem definition, not with a pre-selected AI solution. Evaluate if a simpler, non-AI solution could solve the problem effectively.*

   Rigorous pilot programmes: *Start with small-scale, low-cost Proof of Concept (PoC) projects to test the viability and estimate the true ROI of an AI solution before committing to a large, expensive deployment.*

   Invest in competency: *Build AI literacy across the business, not just in technical teams. When business leaders understand the real capabilities and limitations of AI, they are better equipped to challenge assumptions and select appropriate, cost-effective solutions.*

10. **Failure to start:** Analysis paralysis prevents the business adopting beneficial and low-risk AI tools, causing it to fall behind more agile competitors.

    Suggested risk treatment:

    Identify low-risk, high-value quick wins: *Begin with internal, low-risk applications that boost employee productivity (e.g. internal document summarisation, code generation for developers). Success here builds momentum and confidence for tackling complex projects.*

    Establish a Centre of Excellence (CoE): *Create a small, dedicated team to act as an internal hub for AI experimentation, knowledge sharing, and governance. This provides a safe, controlled environment to explore tools without exposing the entire business to risk.*

    Develop a risk-based approach: *Instead of applying one heavy governance policy to all projects, create a tiered framework. Low risk use cases can proceed with minimal oversight, enabling agility, while high-risk projects are given the rigorous scrutiny they require. This allows the business to move forward responsibly.*

11. **Adversarial attacks and malicious use:** AI models that are not properly secured, can be deliberately manipulated by malicious actors. Techniques like "prompt injection" can trick the AI into bypassing its safety controls, while "data poisoning" can corrupt its training data by intentionally feeding bad data. The goal is to corrupt the model, causing it to fail and it can be used to generate harmful content, extract confidential information, or turn the tool against the business, leading to severe legal, security, and reputational damage.

    Suggested risk treatment:

    Input Sanitisation and Guardrails: *Implement robust filters and validation rules to sanitise all user inputs and model outputs. These "guardrails" should detect and block malicious prompts or harmful generated content in real time.*

    Using a Secure by Design Development Process: *Embed security into the AI development lifecycle. Before building, anticipate likely misuse scenarios and design technical controls to prevent them, rather than trying to add security as an afterthought.*

    Adversarial Testing (Red Teaming): *Proactively test the AI system's defences by simulating attacks. A dedicated "red team" can identify and help patch vulnerabilities before they are discovered and exploited by external threat actors.*

    Strict Access Control and Rate Limiting: *Enforce the principle of least privilege, ensuring AI model only has access to the data and systems it absolutely needs. For public-facing tools, implement rate limiting to defend against automated abuse and denial-of-service attacks.*

12. **Data sovereignty and jurisdictional risk:** Using global public AI platforms can lead to business's data, including sensitive personal information and company confidential information, being stored or processed in overseas jurisdictions. This creates a significant risk of non-compliance with Australian laws (like the Privacy Act 1988) and exposes sensitive data to foreign laws and government access requests, potentially breaching client trust, and legal obligations.

    Suggested risk treatment:

    Ensuring robust data boundaries: *Prioritise and configure AI services that offer explicit regional controls and geofencing. Actively configure the service to lock data processing and storage into a compliant region, such as an Australian-based data centre, to create a robust data boundary.*

    Vendor due diligence and contractual guarantees: *Before procurement, conduct thorough due diligence on the AI vendor's data residency capabilities. Insist on contractual clauses that legally bind the vendor to store and process business owned data only within Australian jurisdictions. Ask the vendor for their data processing map and SOC 2 Type II report.*

Implementing practical controls to detect and prevent data leakage: *Implement appropriate data exfiltration rules to prevent sensitive data being siphoned or transferred out of a sovereign data location. Configure policies to scan for and block outbound emails or prompts that contain patterns of sensitive data (e.g., TFN, Medicare numbers, etc). Configure the firewall or web gateway to block access to common, non-approved public AI tools (e.g., specific chatbot websites) that do not offer Australian data residency.*
Use the identity provider to create a specific user group (e.g., "Approved AI Users") that is the only group permitted to access the sanctioned AI tools. All other employees should be blocked.

Data minimisation and anonymisation: *Where cross-border data transfer is unavoidable, strictly minimise the data sent. Employ anonymisation or pseudonymisation techniques to remove personally identifiable information before the data leaves Australia's jurisdiction, reducing the impact of a potential breach.*

13. **Model theft and extraction:** AI models represent a significant investment in data and computation. Attackers may try to steal a model's intellectual property by repeatedly querying it and using the outputs to reverse-engineer a copy (**model extraction**) to duplicate underlying logic, parameters, and decision boundaries. This constitutes a direct theft of intellectual property, eroding competitive advantage and nullifying significant investment in research, development, and data acquisition.

    Suggested risk treatment:

    Implement robust API rate limiting and monitoring: Enforce strict limits on the number of queries allowed from a single IP address or user account over a given period. Use anomaly detection systems to identify and flag suspicious query patterns that are indicative of an extraction attack, such as an unusually high volume of queries from a new source.

    Output perturbation and confidence score reduction: Introduce a small amount of random noise or rounding to the model's outputs. For classification models, avoid returning highly detailed confidence scores and instead return only the final classification label. This makes it significantly harder for an attacker to accurately map the model's decision boundaries.

    Legal and contractual deterrents: *Include explicit clauses in the API's Terms of Service that prohibit any form of automated querying for the purpose of reverse-engineering, training a competing model, or any other form of model extraction. While not a technical control, this provides a legal basis for action.*

Risk	Suggested Treatment
1. Sensitive data leakage	Implement clear policies, use technical controls (DLP, private sandboxes), and conduct employee training.
2. Data privacy violations	Embed Privacy by Design, confirm legal basis for data use, and enforce data minimisation/anonymisation.
3. Algorithmic bias & discrimination	Audit datasets for bias, conduct fairness testing, use Human-in-the-Loop (HITL), and apply XAI techniques.
4. Disinformation, misinformation & hallucinations	Mandate human verification, use RAG with trusted data, train users on prompt engineering, and use watermarking.
5. Model degradation and drift	Implement continuous performance monitoring, establish a regular retraining cycle, and maintain model versioning.
6. Shadow AI	Use discovery tools (CASB) to find unvetted AI, establish an "Approved List," and enforce centralised governance.
7. Copyright infringement & IP use	Conduct vendor due diligence (indemnification), use "clean" models, and require human review of outputs.
8. Vendor Lock-in	Secure contractual safeguards (data portability), design platform-agnostic architecture, and consider open-source options.
9. Excessive cost & complexity	Adopt a "problem-first" approach, run small pilot programs (PoCs), and invest in business-wide AI competency.
10. Failure to start	Identify low-risk/high-value "quick wins," establish a Centre of Excellence (CoE), and use a tiered, risk-based approach.
11. Adversarial attacks & malicious use	Implement input/output guardrails, follow Secure by Design principles, and conduct adversarial testing (red teaming).
12. Data sovereignty & jurisdictional risk	Enforce regional controls (geofencing), demand contractual guarantees from vendors, and use data minimisation.
13. Model theft and extraction	Implement API rate limiting, use output perturbation (adding noise), and include legal deterrents in terms of service.

Table 1: Summary of AI risks and suggested treatment

8. Risk Management across the AI lifecycle

Effective AI risk management is not a final compliance check; it is a continuous discipline integrated into the entire lifecycle of an AI system. This guide integrates AI risk considerations into a business's governance, ensuring responsible development, deployment, and operation of AI. The following lifecycle stages include questions and considerations to ask at each stage to get the best alignment with the needs of the business.

By embedding governance activities from the earliest strategic discussions to the final decommissioning, businesses can proactively identify and mitigate risks rather than reactively respond to failures. To ensure that risk is considered at every stage, anchor AI governance activities within a familiar project lifecycle.



It is essential that before any AI tools are developed or implemented, the organisational readiness is assessed.

- Consider the roles and responsibilities of individuals in the business, including internal employees and external contractors or vendors.
- Assess existing data assets, data flows, and processes to understand the 'why' behind the existing structure.
- Ascertain the structure and classification of information to understand firstly, how readily available accurate data is to make business decisions, and secondly, to ensure controls are in place to protect privacy and security of information.

Also consider the following questions:

- Will new technology, such as AI, surface information that is obscured but not protected?
- Does the business have an existing responsible use of AI policy? Are any updates required to this?
- Is there clear ownership of business assets and processes? Conduct a RACI to ensure there is clear understanding of roles and responsibilities.

The figures below show the various lifecycle stages in AI Development and are **heavily influenced and based on both ISO/IEC 5338:2023 and ISO/IEC 22989:2022.**
Note that the lifecycle has been extended to reflect the case where the business's role is that of a deployer. As a deployer, one can perform only limited verification and validation.

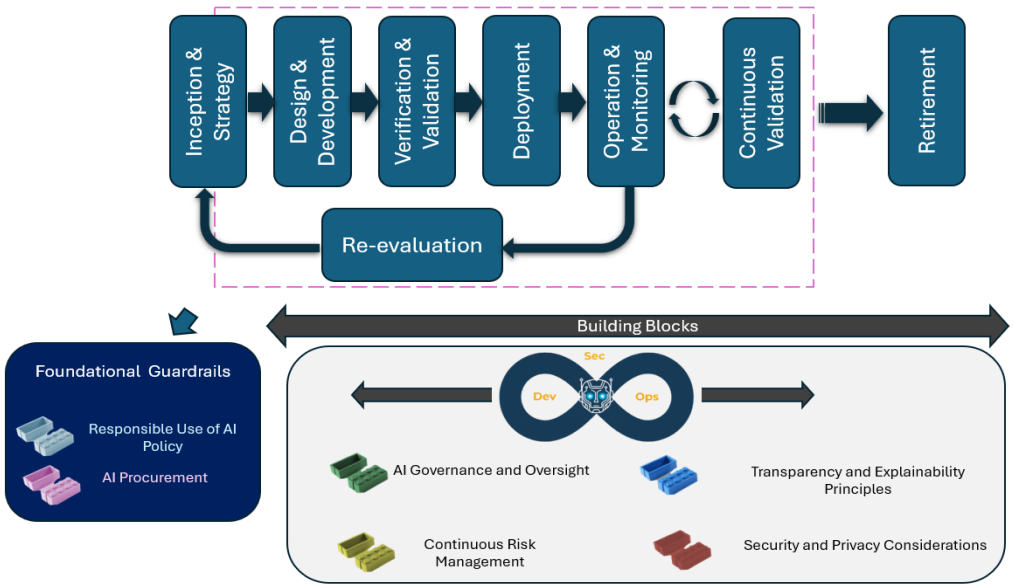


Figure 1: AI lifecycle stages in AI model or AI system development

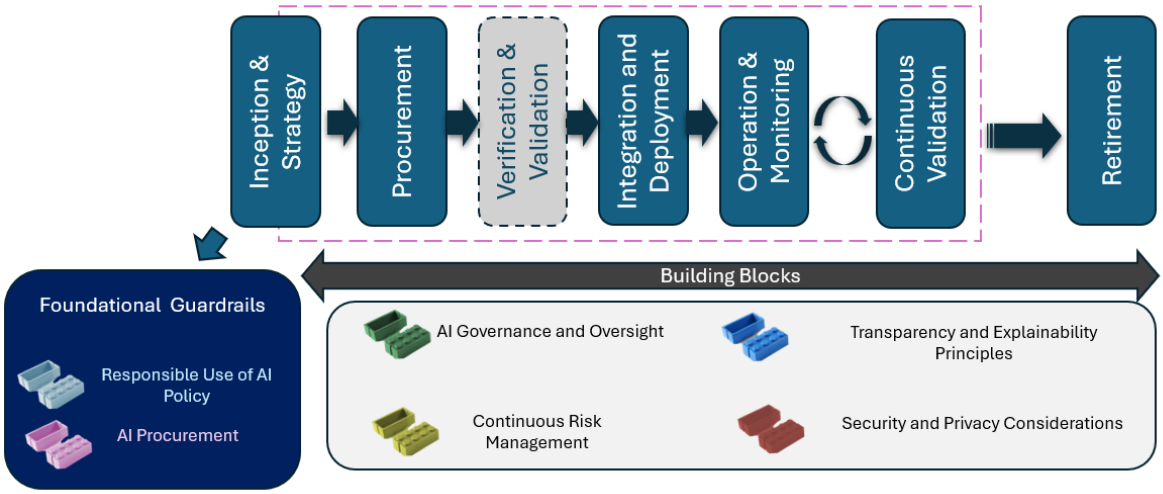


Figure 2: AI lifecycle for AI model or AI system deployment

8.1. Stage 1 - Inception & strategy

This stage is where the planning for AI initiatives, including developing business use cases, are identified and assessed. It is important to not lead with an AI product or vendor but rather

focus on analysing the problem and breaking it down (problem-first, not solution-first). AI may not necessarily be the right solution for every scenario; to consider the benefits, costs, risks, and consequences is an important exercise to ensure that the decision-makers have the facts to make an informed decision.

Consider the following questions:

- Do we understand **why we want to use AI** and what problem it will solve for our business? Is there an alternative available that provides the exact same solution with lesser cost/risks/magnitude of changes required to people, process, and technology?
- What is the role that we are playing? AI Developer, AI Deployer, AI End User?
- Are we involving key stakeholders in the strategy and planning?
- Do we have the right level of executive buy-in?
- Are we aware that using AI still comes with responsibilities, even if we did not build it?
- Do we have a well-defined methodology for AI risk management?
- Is the AI tool/system/ model we are using considered **"high-risk"** (e.g., hiring, credit scoring, biometric surveillance)?
- Have we mapped the **potential ethical, legal, and reputational risks** involved in using this AI?
- Have we established the key principles for responsible and ethical use of AI? Are there **internal policies or guidelines** for using AI responsibly?
- Who will be affected by this project? Consider potential for displaced employees, new people skill requirements, retirement of current processes or assets.
- What are the potential blockers to the success of this AI initiative?

This is the stage to conduct a risk-assessment or a pre-mortem assessment. As part of this, the business's governance approach should align with established local standards.

For instance, ensure that the project's objectives are evaluated against Australia's AI Ethics Principles, such as fairness, privacy protection, reliability, accountability, and transparency, to create a foundation of responsible practice from the very beginning.

Together with key stakeholders, consider the hypothetical scenario of a failed project. What would have occurred to be a failed project? What sequence of events would lead to the failed project? This guide will provide support in such analysis.
As part of this, read all the following lifecycle stages, and next, consider the successful project outcome juxtaposed to the hypothetical failed project. Performing this exercise will provide a solid basis from which to conduct risk management.

8.2. Stage 2 - Design & development, or procurement, depending on the role

The purpose of this phase is to determine how the solution will be built or procured to achieve the business's objectives.

The challenge in this stage is ensuring that business value will be delivered through the solution, whilst ensuring security, privacy and ethical considerations have been applied.

In the world of AI, it is important to understand to what extent AI will be present in all projects. Even a *"non-AI project"* may still use AI during its development. Consider the following scenarios where AI may be present:

- Development of an AI system, including fine tuning of existing AI models
- A Software-as-a-Service (SaaS) solution that may use AI behind the scenes
- Development conducted by internal developers or external vendor, where generative AI tools may be used to generate parts of the final solution

Consider the following questions when selecting a vendor:

- Does the AI vendor perform threat modelling specifically for the AI models and the systems they are integrated into? Can they describe the methodology?
- How do they account for AI-specific threats such as model inversion, data poisoning, adversarial attacks, and model extraction in the threat models?
- How does the AI vendor test for and remediate vulnerabilities in the AI model's software supply chain, including third-party libraries and dependencies?
- Can the AI vendor provide evidence of their security certifications and attestations? Can they share their SOC2 Type II reports that provide independent, third-party assurance about the operational effectiveness of a vendor's security controls over a period, usually six or twelve months.
- How frequently does the AI vendor conduct penetration testing and vulnerability assessments on their AI systems? Are these performed by independent third parties?
- Does the AI vendor provide clear documentation on what their AI system does and how it works? Is this in a language that is easily understandable, or does it contain technical jargon? Note that transparency is a good indicator of ethical maturity.
- Can the vendor explain how the AI system was trained — including where the data came from? Consider things like ethical sourcing, respect for copyright, consider data provenance and lineage (origin of data, source of data, what transformations were applied). Do they ensure the accuracy, transparency, and ownership of data in scope?
- Does the vendor provide clear terms and conditions for if and how the business's data may be used (seek legal advice), including for training the model. Does the agreement specifically address 'algorithmic disgorgement,' and outline the process for the businesses to demand that their data's influence is removed from the model? Consider a trustworthy and reliable vendor that is transparent and upfront about how it will use the business's data for example whether it will anonymise, aggregated data, or potentially provides an option for data/ model destruction (this is not always feasible, so understand the implications around sharing data).
- Who owns the IP, training data, and user data? Is this contractually clear?
- How does the AI vendor ensure the segregation of the business's data from that of other customers, both in training and in production?
- Where will the business's data be processed and stored? Will cross-border data transfers be managed in compliance with regulations?

- What are the AI vendor's data retention and deletion policies for business inputs and the AI-generated outputs? Who has access to the business's data, and what access control measures are in place?
- Has the vendor assessed the bias and fairness of their model? How is it monitored post deployment?
- Can the vendor demonstrate compliance with the Privacy legislation, or other applicable laws (e.g., the Australian Privacy Act 1988, European Union's AI Act, GDPR)?
- Are there terms of use or ethical commitments the vendor follows (e.g. codes of conduct, model cards, impact assessments)?
- What aspects of the AI's behaviour can users customise?
- Is responsibility or shared responsibility clearly understood?
- What is the vendor's policy on using customer data for training or fine-tuning AI models? Is this an opt-in or opt-out process?
- How agile is the vendor in terms of choosing a model? Is it easy to switch to a different model to reduce cost or improve accuracy and performance?

8.3. Stage 3 - Verification & validation

The purpose of the **verification** process is "to provide objective evidence that a system or system element fulfils its specified requirements and characteristics".

As ISO/IEC 5338 explains, verification is achieved statistically by applying the model to test datasets and measuring its performance against the quantitative metrics defined in the system requirements (Stage 2). This includes metrics for correctness (e.g. accuracy, precision, recall) and robustness.

The **validation** process aims for "providing objective evidence that the system, when in use, fulfils its business or mission objectives and stakeholder requirements" in its intended operational environment.

The following should be considered:

- Performing bias audits on the model's predictions across different demographic subgroups using the predefined fairness metrics.
- Conducting "Stress Testing" by evaluating the model's performance on known edge cases, noisy data, and adversarial examples to understand its failure modes.
- Testing the model for effectiveness, accuracy, and efficiency. It should also address the inclusive, fair, and transparent completion of tasks.
- Actively attempting to compromise the model through adversarial testing techniques to measure its resilience against evasion or data poisoning attacks.

The goal is to validate that the system is not only technically sound but also usable, valuable, and trustworthy to the people it is intended to serve.

8.4. Stage 4 - Deployment

The purpose of this phase is to ensure successful implementation of the solution. It is not just a technical installation but must consider making the organisational change necessary to embed the solution. Consider the Unfreeze, Change, Refreeze model by Kurt Lewin (see https://www.prosci.com/blog/lewins-change-theory):

- **Unfreeze**: Prepare the business for change. Ensure all stakeholders understand what the change is, why it is necessary, and what the business will look like after the change.
- **Change**: Implement the changes, including people, process, and technology changes.
- **Refreeze**: Once the change has been implemented, focus on supporting the business and embedding the change, reinforcing the new status quo.

Please note that it is important to have a rollback plan in place when deploying an AI system. Deployment may not always be successful, and unintended consequences can arise post-launch, such as unexpected errors, biased outputs, or security vulnerabilities. Having contingencies ready allows the business to quickly revert to a previous stable state and minimize operational disruption and reputational damage.

Prior to deployment, consider the following questions:

- Have we clearly defined how the AI system will be used in our business processes?
- How will the “handover” to business-as-usual take place? Is the ownership clearly understood and accepted as part of the “handover”?
- Will humans still be **involved in making key decisions**, especially if the AI output affects people?
- Accuracy - Have we tested the outputs from the AI system for hallucination?
- Security
    - Has red teaming and penetration testing been performed?
    - Are there input prompt guardrails in place to prevent use of profanity, abuse, malicious use, or leak of sensitive or intellectual information?
    - Are there output guardrails to prevent leak of organisational source code or sensitive/competitive information?
- Transparency
    - Are we informing customers or employees that AI is being used?
- Are we being transparent (there may be obligations around the same, especially if interacting with EU citizens). Before the system goes live, a clear communication plan must be executed. Note: this is also part of "System Disclosure" principle and is a legal requirement under regulations like the EU AI Act, which mandates transparency when individuals interact with certain AI systems.
    - Do we have the right disclaimers in place?
    - Does right to challenge the decisions exist, and do we allow talking to a human?
- Resilience - Can the solution withstand and recover from disruptions and failures?

- Explainability - The lifecycle stage must clearly articulate the solution's purpose, functionality, restricted user groups, unsupported user groups, and any known limitations.
    - Are users clear about what the system can and cannot do?
    - Is it possible to explain how and why the system makes specific decisions?

  If they are not trained on its limitations, they may over-rely on it and make critical errors.
- Accountability
    - Have we assigned a person or team to be **responsible** for overseeing the AI system? Is there a human in the loop?
    - Are there ways to **override** or challenge the AI system's decisions meaningfully?
- Do the designated human reviewers have the necessary tools, training, and authority to intervene, and is there a clearly defined, accessible, and functional appeal process for affected individuals?
- Have we tested the AI tool on a **small sample or pilot** before rolling it out broadly?
- Auditability - Can we track and review the AI solution's processes and decisions?
- Accountability - do we have someone accountable for the AI solution's decisions?
- Contestability - Are there mechanisms for users to challenge and appeal decisions made by the AI solution.

8.5. Stage 5 - Operation & monitoring

AI, unlike traditional software evolves over time with use. Hence, any operation must include periodic monitoring, with more rigour if the AI use case is high risk. As a Developer, design for re-evaluation and continuous validation of the AI model, and if only the Deployer, ensure communication channels with the Developer of the model to help any challenges with AI model performance. This phase involves embedding new ways of working and can be considered the 'refreeze' component of the Unfreeze, Change, Refreeze model by Kurt Lewin referenced above.

This phase should also include a project retrospective activity to ensure the project team assesses both the project and the delivered solution. Is the delivered solution successfully adding business value?

Consider whether solution delivers its services, sustain the capability of the system to continue delivery of the services, include having an active program of oversight to maintain its performance, integrity, fairness, and safety over time. Also, consider common corrective actions triggered by performance degradation include model retraining or re-engineering to address "data drift" (change in the statistical properties of the input data used to train a model) and "concept drift" (changes in the underlying relationships the model has learned).

Also, consider the following questions:

- How will we collect feedback?

- Have we integrated AI incident response management with the overall incident management?
- Are we tracking how well the AI system is performing over time?
- Are there **unexpected behaviours**, errors, or unfair results we need to address?
- Can users **report problems or appeal decisions** made or suggested by the AI system?
- Do we know and are okay with the kind of **data the AI tool collects from us** during use?
- What will be the triggers for model retraining?
- Is the system **getting updated**, and do we know what changes have been made?
- Does the model/ system have Model Rollback Capability?
    - This is a critical safety mechanism to quickly revert to a previous, known-good version of a model if a newly deployed model exhibits severe performance issues or unexpected behaviour. This rollback capability must be tested and ready for immediate activation.
- Ongoing Data Governance
    - Are we complying with data protection obligations (e.g., informed consent, data minimisation, data subject rights)?
    - How are the developers performing ongoing monitoring of privacy and compliance to ensure the system continues to adhere to data minimisation principles and consent agreements?
- Are the operating costs tracking as expected?
- Are there any questions or concerns from the new owner following “handover”?

Additionally, it is strongly recommended to schedule periodic external reviews or audits, particularly for high-risk AI applications. These independent assessments help verify compliance, fairness, security, and ethical use, providing assurance that the AI system continues meeting business and regulatory standards.

8.6. Stage 6 - Retirement

This phase is focused on ensuring there is a process for managing the retirement of AI systems, including any associated assets, data models or other resources explicitly associated with the AI system. At retirement, these systems should be decommissioned, and data should be appropriately migrated or destroyed. The way in which this is executed should be formally planned, secure and responsible.

Consider the following questions:

- Do we have a plan to **stop using the AI system** if it becomes non-compliant, harmful, or irrelevant or no longer deliver the planned benefits i.e. ROI?
- Do we have a disposal plan for the model and/or data? Consider
    - The production models themselves.
    - All training, validation, and test datasets, which may contain sensitive or personal information.
    - User data collected by the system during its operation.
    - The source code and any supporting infrastructure.

- Can we retrieve our data, AI model/algorithms, and IP from the AI system if we switch vendors? Note that this must be built into the initial acquisition agreement during Stage 2, with clear specifications for data formats, retrieval mechanisms, and associated costs.
- Have we learned anything from using AI that can **improve our risk processes** next time?
- Do we have a documented Business Continuity Plan (BCP) and fallback strategy if the AI system fails or must be shut down *instantly*? If we move back to manual processes, are there documented manual processes and employees who are regularly trained to perform them?

Lifecycle Stage	Key Risks to Consider
Inception & strategy	• Excessive cost & complexity • Failure to start • Algorithmic bias & discrimination • Data privacy violations • Data sovereignty and jurisdictional risk • Shadow AI
Design & development / Procurement	• Vendor Lock-in • Copyright Infringement and Unauthorised Use of Intellectual Property (IP) • Data sovereignty and jurisdictional risk • Adversarial attacks and malicious use • Model Theft and Extraction • Algorithmic bias & discrimination • Sensitive data leakage • Data privacy violations • Misinformation, disinformation and hallucinations
Verification & validation	• Algorithmic bias & discrimination • Adversarial attacks and malicious use • Misinformation, disinformation and hallucinations • Model degradation and drift
Deployment	• Misinformation, disinformation and hallucinations • Adversarial attacks and malicious use • Sensitive data leakage • Algorithmic bias & discrimination • Data privacy violations
Operation & monitoring	• Model degradation and drift • Shadow AI • Algorithmic bias & discrimination • Adversarial attacks and malicious use • Model Theft and Extraction • Misinformation, disinformation and hallucinations • Sensitive data leakage • Data privacy violations
Retirement	• Vendor Lock-in • Sensitive data leakage • Data privacy violations • Copyright Infringement and Unauthorised Use of Intellectual Property (IP)

Table 2: AI lifecycle stages and key AI risks

9. Establishing accountability and an action plan

With a clear understanding of the risks, a lifecycle framework in place, and a plan for cultural development, the final step is to operationalise governance. This requires establishing clear lines of accountability and defining a concrete action plan to translate strategy into practice. This section provides the essential tools for Senior Management or The Leader to implement a successful AI programme.

9.1. The AI governance RACI matrix

Ambiguity in roles and responsibilities is a significant threat to any cross-functional initiative, and AI governance is no exception. A RACI (Responsible, Accountable, Consulted, Informed) matrix is a standard governance tool that is indispensable for clarifying ownership. It forces critical conversations and ensures that for every key AI-related task, there is a single point of accountability.

Whilst SMEs often have individuals wearing multiple hats, the exercise of creating a RACI chart is crucial for ensuring no critical responsibilities fall through the cracks.

The following matrix provides a foundational template for assigning AI governance responsibilities. SMEs should adapt this to their specific organisational structure.

Note that dedicated roles like a Chief Risk Officer (CRO) or General Counsel may not exist in all organisations. In such cases, these responsibilities should be mapped to the existing structure; for example, the CRO function may be handled by the CFO, and legal counsel may be an external advisor. Sometimes, there are no dedicated CTO or even CISO roles, depending on the size and maturity of the business. In such a scenario, the appropriate individual, whether the CIO or their appointed delegate, can take up the role.



The key is that there is someone in the business who is assigned the responsibility, not necessarily the title. This avoids things "falling through the cracks" as no one was assigned the responsibility, and there was a tacit assumption that "someone" would take care.

The following matrix provides a foundational template for assigning AI governance responsibilities. SMEs should adapt this to their specific organisational structure.

Responsibilities	Senior Management	CTO	General Counsel/ Legal	Chief Risk Officer (CRO)	CISO/ Cyber Security and Data Governance	Business Team	Solution Support Team
Define AI Solution Requirements	A	I	I	I	I	R	C
Approve AI Vendor & Contract	A	R	R	C	C	R	C
Develop AI Acceptable Use Policy	I	C	R	A	R	C	C
Complete Security & Privacy Assessment	I	I	C	C	A/R	I	C
Manage & Monitor AI Operation	I	R	I	C	I	A	R
Perform Bias & Fairness Audits	I	C	C	A	I	R	R
Manage AI-Related Incidents	I	R	C	C	R	A	R
Monitor for New AI Laws & Regulations	I	I	A	R	I	I	I
Approve AI System Decommissioning	A	R	I	I	C	R	C

(R=Responsible, A=Accountable, C=Consulted, I=Informed)

Table 3: AI Governance RACI Matrix

10. Conclusion

In an environment where competitors are rapidly adopting artificial intelligence, literacy and innovation have become prerequisites for survival. A perilous gap has, however, opened between expectation and reality. While nearly 90 percent of leaders anticipate that deploying AI will drive revenue growth, the reality is 95% of AI projects fail to deliver measurable returns (according to the Massachusetts Institute of Technology (MIT)), and nearly 70 percent of all major corporate transformations fail[$].

The critical task for The Leader, therefore, is to navigate this gap. Success hinges not on the impossible goal of eliminating risk, but on the strategic discipline of managing it effectively. For The Leader working at the SMEs, this guide offers that framework, demonstrating that robust governance is not an obstacle to progress but the very foundation upon which lasting innovation is built. By embracing the dual roles of Translator to articulate technical threats as clear business risks; and Architect to design responsible innovation from the ground up, The Leaders can support CEOs in guiding their companies with greater clarity and confidence.

Moving forward successfully requires a proactive approach, integrating risk management directly into the entire AI lifecycle. The effort must begin with a solid data foundation, because the most critical risks—algorithmic bias, data privacy breaches, or IP theft—almost always emerge from weaknesses in data governance. For most SMEs, this means adopting the role of a responsible 'Deployer', a position that demands rigorous vendor due-diligence and an insistence on transparency through instruments like the AI Bill of Materials (AI BOM).
(See A3.2 The AI Bill of Materials (AI BOM) for details).

However, technical safeguards alone are not enough and without addressing human factors, transformations will not be successful. Sustainable governance is built upon three core organisational commitments.

- A culture of responsibility that is actively championed by Senior Management
- Clear accountability, using a RACI matrix to define ownership precisely.
- A practical roadmap for implementation, like the 90-day action plan provided in the appendix.

AI governance is the strategic enabler that allows leaders to move beyond 'analysis paralysis' and do more than just mitigate risk. It is the key to building a resilient, trustworthy AI program that unlocks genuine innovation, drives competitive advantage, and secures the organisation's future in an AI-powered world. The practical examples provided in the appendix offer a safe, managed way to begin this journey.

With employees prepared to adopt artificial intelligence, it is now incumbent upon management to provide clear direction and strategy.

$. “Almost all companies invest in AI, but just 1 percent believe they are at maturity”— McKinsey research

Appendix:

A1. The First 90 Days in practice: The Leader's handbook

*This is a sample actionable playbook that assumes that the organisational role is that of a **Deployer** and provides an indicative timeline with recommended activities for The Leader to get AI initiative functioning as quickly as possible in a controlled and responsible manner. It provides an indicative timeline and clear, prioritised activities to accelerate the launch of AI initiatives while embedding strong governance from the outset.*

Month 1: Assess & discover (Days 1-30)

- **Form a cross-functional AI governance task force** comprising stakeholders from IT, security, legal, HR, and key business units. This team will champion the governance initiative, maintain and own the AI risk register and oversee policy implementation.
- **Conduct an AI usage inventory** to identify and catalogue all known AI systems currently in use, both sanctioned enterprise tools and known instances of "Shadow AI."
- **Perform an initial high-level risk assessment** using the STRIDE/ DREAD-AS (see appendix) framework, conduct a preliminary assessment of the top 3-5 AI systems identified in the inventory to establish a baseline risk posture.
- **Brief the Senior Management** on the initial findings and develop the business case for the governance initiative to secure the mandate and resources for ongoing effort.

Month 2: Build the Foundation (Days 31-60)

- **Conduct foundational data planning**
    - Based on the high-risk systems found in Month 1, identify the minimum datasets required.
    - Partner with data owners to ensure this data has clear classification, ownership, and tags (e.g., 'Confidential', 'Internal' etc).
    - Assess source system compatibility and flag any "black box" legacy systems that pose a risk.
- **Develop version 0.1 of the AI AUP** focusing on clarity, acceptable use cases, and the most critical prohibitions (e.g., leaking sensitive data). Communicate it across the business. This is the foundational governance document. It must be clear, simple, and enforceable.
- **Develop the AI vendor due diligence checklist** to evaluate the top three to five critical AI vendors, supporting secure procurement and ongoing vendor management.
- **Launch mandatory foundational AI literacy training** Roll out an initial, mandatory training module for all employees covering the basics of AI, the key risks (hallucinations, bias, data privacy), and the newly established AUP.

- **Identify departmental AI champions,** working with department heads to nominate candidates who will promote responsible AI withing their respective areas.

Month 3: Implement & enforce (Days 61-90)

- **Implement technical security guardrails** to detect and prevent the exposure of proprietary code or sensitive information.
    - **Block unapproved AI sites** and block access to high-risk, unapproved AI tools via secured web gateways or firewalls.
    - **Deploy Data Loss Prevention (DLP)** by configuring DLP rules to detect and block the pasting of sensitive data patterns (PII, financial data, project codenames)
    - **Enforce secure access** and use SSO and MFA for all approved enterprise AI tools.
    - **Enable logging** and ensure all approved tools have comprehensive logging enabled to track usage and provide an audit trail.
    - **Create AI specific user group** (e.g., "Approved AI Users") using the business's identity provider. This should be the only group permitted to access the sanctioned AI tools. All other employees are blocked.
- **Formalise the AI champion network** and convene the meetings to track progress on AI adoption with the identified AI champions. Provide them with advanced training and establishing clear collaboration and escalation channels.
- **Present the complete AI governance framework**, including the risk assessment findings and RACI chart, along with the AI roadmap to Senior Management for endorsement and to secure ongoing support.
- **Schedule review cycle for the AI governance task force** to reassess tools, evolving risks, and the effectiveness of existing governance controls.

A2. Beyond the 90 days- Ongoing AI Solution Considerations

- **Monitor AI performance**

  Continue monitoring AI system performance, focusing on technical outcomes, regulatory compliance, and risk management. Ensure ongoing data governance to meet data protection obligations, monitor privacy, and safeguard data residency and jurisdictional requirements. Some questions to consider are:
  Are we tracking how well the AI system is performing over time?
    - Are we performing ongoing data governance?
        - Are we complying with data protection obligations (e.g., informed consent, data minimisation, data subject rights)?
        - Understanding data compliance requirements, data residency, and jurisdictional concerns

        - How are the developers performing ongoing monitoring of privacy and compliance to ensure the system continues to adhere to data minimisation principles and consent agreements?
    - Are there unexpected behaviours, errors, or unfair results we need to address?
    - Can users report problems or appeal decisions made or suggested by the AI system?
    - Do we know what kind of data the AI tool collects from us during use? Are we comfortable with that?
    - What will be the triggers for model retraining?
    - Is the system getting updated, and do we know what changes have been made?
    - Does the model/ system have model rollback capability?

- **Offboarding- secure decommissioning / retirement**
  When decommissioning AI systems, have a secure retirement plan that includes certified data destruction, contract termination processes, and structured data retrieval from vendors. Capture learnings from each deployment to continuously refine organisational risk management strategies.
    - Do we have a plan to stop using the AI system if it becomes non-compliant, harmful, or irrelevant?
    - Do we have a disposal plan for the model and/or data? Is there a process for the certified destruction of our data upon contract termination?
    - Can we retrieve our data from the AI system if we switch vendors?
    - Have we learned anything from using AI that can improve our risk processes next time?

A3. AI procurement and securing the AI supply chain

For organisations acting as AI deployers, procurement represents the most critical control point for managing third-party AI risks. The focus is not on training models from scratch but on the rigorous selection, contracting, and integration of third-party AI systems. Failure to conduct thorough, AI-specific due diligence at this stage means inheriting unquantified and potentially unmanageable risks from vendors.

A3.1 AI-Specific vendor due diligence

Standard vendor security questionnaires and checklists are fundamentally inadequate for assessing the unique risks posed by AI systems. The due diligence process must be expanded to

probe the vendor's own AI governance and risk management practices throughout their development lifecycle.

Below is a due diligence framework, (which can serve as a vendor checklist), covering the essential domains of inquiry.

Governance, risk, and compliance considerations

This area assesses the vendor's organisational maturity and commitment to managing AI risk.

AI governance: Request evidence of a formal AI governance program, such as an AI ethics policy or certification (e.g., ISO 42001).

Regulatory compliance: The vendor must attest that the system complies with all relevant laws for the jurisdictions in which you operate (e.g., Privacy Act, EU AI Act).

Performance and reliability: What are the terms in contractually binding Service-Level Agreements (SLAs)? If the AI model fails, what is the fallback behaviour? Does it degrade gracefully (e.g., to a simpler rules-based system) or fail completely?

Model lifecycle management: What mechanisms are in place to monitor for model drift? How are clients alerted if the model's accuracy or performance degrades?

Supply chain diligence: The vendor must demonstrate their own due diligence process for all components, data, and services used to build and train their AI system.

Data provenance, handling, and privacy considerations

This is critical for protecting your proprietary information and avoiding legal risk.

Data provenance: What specific datasets were used for training? The vendor must provide evidence of data legality, showing it was sourced ethically and without infringing on copyright or privacy rights.

Data usage and confidentiality: Is there a contractual guarantee that your business data (prompts and outputs) will not be used to train or fine-tune models for other customers? This is often a non-negotiable control, especially if you are concerned about intellectual property.

Data sovereignty: Where geographically will your data (inputs, outputs, metadata) be stored and processed?

Data segregation: How is your data logically and/or physically segregated from other tenants? What controls prevent your data from 'leaking' into another customer's outputs?

Privacy controls: Does the vendor provide mechanisms to detect, mask, or anonymise personal information (PI) before it is processed by the model?

Data ownership: The contract must clearly state who owns the inputs provided and, critically, who owns the outputs generated by the model.

Transparency, fairness, and explainability considerations

Vendors must provide an AI Bill of Materials (AI BOM), detailing training data sources, model types, software dependencies, and steps taken to test for and mitigate algorithmic bias in training data and outputs.

Bias and fairness testing: Vendors must move beyond vague claims of "ethical AI." Request formal reports from their internal or third-party fairness audits. These reports must specify the fairness metrics used, the demographic groups tested (e.g., racial, gender, age), and the mitigation steps taken.

Explainability (XAI): For high-risk decisions (e.g., a risk score, a candidate ranking), can the model explain why it reached a conclusion? Vendors should offer tools enabling interpretation of outputs and conduct independent fairness monitoring.

Intended use and limitations: The vendor must clearly document the model's intended use cases and, just as importantly, its known limitations and inappropriate uses. Deploying a model for a purpose it was not designed for introduces unacceptable risk.

Security and incident response

Vendors of high-risk AI systems should provide attestations or third-party audit reports demonstrating compliance with relevant standards and laws, such as the EU AI Act or sector-specific regulations. Vendors must also perform due diligence on their own supply chains, including third-party risk management processes for all components, data, and services used in AI development.

Model security: How are the vendor’s AI models protected from data poisoning, model evasion, and other adversarial attacks? Request results from penetration tests or other security validation.

Cybersecurity posture: Can the vendor provide a recent SOC 2 Type II report or equivalent third-party security audit?

AI-specific incident response: Does the vendor have a documented incident response plan that explicitly covers AI-related security breaches? What are the specific, contractually binding timelines for data breach notification involving the customer’s data?

A3.2 The AI Bill of Materials (AI BOM)

An AI Bill of Materials (AI BOM) is a non-negotiable component of procurement for significant AI systems. It is a structured inventory of all AI system components, analogous to a Software Bill of Materials (SBOM), providing foundational transparency for risk management and governance.

An effective AI BOM must include:

Datasets: A list of all training, validation, and testing datasets, including their provenance (origin and data lineage) and any preprocessing steps.

AI models and architecture: Detailed information on the model architecture (e.g., transformer), algorithms, key hyperparameters, and versioning.

Software components: An inventory of all third-party software libraries and frameworks, crucial for vulnerability management.

Hardware requirements: Specifications for computational infrastructure, including GPUs, cloud instance types, and networking hardware.

Security and governance: A summary of security measures (data encryption methods, access control) and results of fairness and bias testing.

Ownership and processes: Identification of the model developers, owners, and the operational workflows (e.g., CI/CD pipelines) used in its creation.

Third-party supply chain management: Risk management processes for vendors and suppliers throughout the supply chain.

The AI BOM is far more than a procurement checklist; it is a foundational artifact that serves as a cornerstone for effective risk management across all subsequent stages of the AI lifecycle.

Its absence creates dangerous blind spots and introduces significant operational friction downstream. For example, during the Design and Integration stage, the security team cannot architect appropriate guardrails or proactively manage vulnerabilities without knowing the specific software libraries and dependencies listed in the AI BOM.

In the Testing and Validation stage, an AI red team cannot effectively model threats without understanding the model's architecture and data sources, and fairness audits are rendered meaningless without the training data demographics documented in the BOM.

During Deployment, crafting accurate transparency notices for users is impossible without the model's capabilities and limitations, which are core components of the AI BOM's metadata.

Finally, in the Operation and Monitoring stage, diagnosing model drift requires the baseline data distribution information from the BOM, and an incident response team cannot determine the root cause of a failure without the complete system inventory it provides.

Therefore, mandating a comprehensive AI BOM during procurement is a powerful, proactive risk management strategy. It shifts the burden of transparency from the deployer to the developer and provides the essential ground truth required for effective security, testing, transparency, and operational monitoring throughout the system's life.

A3.3 Contractual safeguards and liability allocation

The procurement contract is the primary legal mechanism for risk allocation and vendor accountability. Negotiate AI-specific contractual safeguards, including:

Data usage prohibition: The contract must contain an unambiguous clause that explicitly prohibits the vendor from using the business's data including inputs (prompts) and outputs to train or improve their general-purpose AI models.

This should be a clear "no" in writing, with no exceptions for anonymised or aggregated data *unless explicitly approved* after a thorough risk assessment. This safeguard is essential to prevent the leakage of proprietary information, trade secrets, or sensitive customer data into a model shared by other customers.

Liability and indemnification: Clear allocation of liability for harms caused by the AI. Will the vendor indemnify you (cover legal costs) if you are sued for copyright infringement, discrimination, or other harm caused by using the AI's output? Vendor attempts to disclaim all liability must be challenged.

Audit and monitoring rights: A clause providing you the right to audit the vendor's controls or, at minimum, receive regular third-party audit reports (e.g., SOC 2).

Data portability and certified deletion: A clause guaranteeing your right to retrieve all your data in a usable, structured format upon termination. It must also obligate the vendor to provide certified destruction of all your data from their systems.

This is a critical control for the final stage of the AI lifecycle.

Compliance with laws: Require guarantees that the vendor's system complies with relevant laws (e.g., Privacy Act, EU AI Act), ensuring that personal information generation is reasonably necessary, lawful, and fair.

SLAs and penalty clauses: These must be clear and measurable, establishing specific financial or service credit consequences for failing to meet standards for performance, security, or compliance.

By consolidating vendor evaluation, transparency requirements, and contractual protections into a unified framework, technical leaders can streamline AI procurement while maintaining rigorous control over third-party risks.

A4. Deployer vs Developer considerations

SMEs can proceed with the use of AI using one of the following two approaches:

1. Adopt a commercial product, assuming the role of AI Deployer (more likely)
2. Develop, fine-tune, retrain or extend AI functionality, assuming the role of AI Developer

As an AI Deployer, businesses must conduct due diligence prior to adopting a commercially available AI solution to ensure it is safe, compliant, and genuinely fit for purpose. This process protects the business from financial, reputational, and legal risks. Securing the AI procurement

and supply chain becomes even more important, due to less control and more dependencies on third parties. Also, the obligations around understanding AI's business objectives and use cases is important, as is the obligation to provide users with transparency, and practice responsible deployment.



EU Artificial Intelligence Act

As per EU AI Act, Developer aka 'provider' means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge.

But put simply, these are the businesses that develop an AI System or GPAI model for commercial purposes including internal usage. The provider is the creator of the AI system. They are responsible for the initial design, development, and ensuring the AI system complies with the regulations.

For deployers, EU AI Act 'deployer' means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity. Or simply, deployers are organisations that use an AI system under their authority and deploy it in their organisation in a professional context.

Key areas of due diligence should include:

- Vendor due diligence (see AI procurement and securing the AI supply chain)
- Prior testing for the intended application
  - Evidence of performance
    - What verifiable evidence can the vendor provide of the product's performance in a similar industry or for an identical application? Request case studies, independent reviews, or performance benchmarks.
  - Contextual robustness
    - How does the model perform when faced with data or scenarios specific to your business or Australian context (e.g., local terminology, demographics, or regulations)?
  - Proof of Concept (PoC)
    - Is it possible to conduct a trial or PoC using the business's own anonymised data in a secure, sandboxed environment? This is the most effective way to validate a vendor's claims.
  - Edge cases and limitations
    - What are the known limitations or "edge cases" where the product is known to perform poorly? An honest vendor should be transparent about these.
  - Ongoing accuracy

        - How does the product handle "data drift" (when the input data changes over time, potentially reducing model accuracy)? Does the vendor have a process for monitoring and retraining the model?
- The integration of human oversight into processes
    - Meaningful intervention points
        - At what specific points in the process will a human review, validate, or override the AI's output? How is this "human-in-the-loop" mechanism designed?
    - Decision-making support
        - Does the product's interface provide enough context or explanation for the human overseer to make an informed judgement? Can they understand why the AI reached a particular conclusion?
    - Training and skills
        - What specific training and skills will your employees need to effectively manage and oversee the tool?
    - Escalation pathways
        - What is the formal process for handling disagreements between the human expert and the AI, or for when the AI's output is clearly erroneous or harmful?
    - Audit trails
        - Are all human oversight actions (e.g., approvals, overrides, corrections) logged in an auditable trail? This is vital for accountability and continuous improvement.
- Potential privacy and security vulnerabilities
    - Data handling and privacy
        - Does the product process personal information? If so, where is the data stored and processed (i.e. data residency)? This is critical for compliance with the Australian Privacy Principles (APPs)
        - How is data protected both in transit (e.g. TLS 1.3 encryption) and at rest (e.g. AES-256 encryption)?
        - Does the vendor's privacy policy align with obligations under Australian law?
        - Separation of data between customers and clarity on whether the customer data may be used for training the model and the scope of such usage
    - Security posture
        - Has the vendor undergone independent security assessments, such as penetration testing or a SOC 2 audit? Can they provide certification reports (e.g. ISO 27001)?
        - How will the product integrate with your existing security tools, such as single sign-on (SSO)?
        - What is the vendor's documented process for identifying and patching security vulnerabilities, and how will they notify the business of an incident?

- Protocols governing access to any personal information the product uses or generates
    - Vendor access
        - Who within the vendor's organisation (e.g., support engineers, data scientists) can access the business’s data, and under what circumstances? Are these access rights strictly controlled and logged?
    - Customer-side controls

        - Is it possible, as the customer, to define granular roles and permissions to ensure employees can only access the data and features necessary for their jobs?
    - Data ownership and IP
        - Does the contract clearly stipulate data ownership and retention policies around business's data? What about any new data or intellectual property generated by the business's use of the product?
    - Data deletion on offboarding
        - What is the process for retrieving the business's data and ensuring it is securely and permanently deleted from the vendor's systems once the contract ends or is terminated? The contract should guarantee a "right to be forgotten."

On the other hand, looking to develop, fine tune, or retrain their own AI models, the obligations and responsibilities are significant, yet different involving more technical knowledge and model considerations like feature engineering, model selection, training. A pragmatic and risk-based approach is crucial. The investment in governance and safety should be proportionate to the potential impact of the AI model.

Some of the activities and responsibilities that demonstrate due diligence by AI Developers include

- Foundational governance and risk management
    - 'Privacy by Design'
        - Privacy should be a core component from the very beginning, not an afterthought. This involves conducting a Privacy Impact Assessment (PIA) before development starts, to identify and mitigate any risks to the handling of personal information, in line with Australia's Privacy Act.
    - Conduct comprehensive risk assessments
        - Go beyond privacy. Evaluate and document potential risks across various domains:

          Data security: Unauthorised access, data breaches.

          Operational: Model failure, incorrect outputs.

          Ethical & reputational: Unfair bias, lack of transparency, public backlash.

          Legal & compliance: Breaches of industry standards or laws.
    - Maintain robust documentation

      Diligent record-keeping is critical for accountability, debugging, and demonstrating due diligence. This should include:

        - Data provenance and lineage: Where the training data came from, what transformations have been applied, its characteristics, and limitations.
        - Testing results: Records of performance, bias, and stress tests.
        - Vendor due diligence: If using third-party tools or platforms, a standard IT vendor checklist is insufficient. Your due diligence must be expanded to cover the unique, complex risks introduced by AI.
        - Human oversight logs: Records of when and why human intervention was required.
        - Model versioning: Clear records of changes between different model iterations.

- Data and Model Integrity

  The quality and fairness of the model depends entirely on the data on which it is trained.

  - Audit datasets for bias and quality
    - Before training or finetuning, thoroughly analyse your datasets. Ensure the data is accurate, relevant, and representative of the environment in which the model will operate. Actively search for and mitigate potential biases related to demographics, location, or other attributes to prevent discriminatory outcomes.
  - Evaluate models using fairness metrics
    - Do not rely solely on accuracy as a measure of success. Use established fairness metrics to check if the model performs equitably across different subgroups. Document these findings as part of the model's testing results.

- Technical security and control

  Implement technical measures to ensure the model operates safely and as intended.

  - Implement technical guardrails
    - These are security controls built into the system to enforce policies. Examples include strict access controls for who can retrain or deploy models, data encryption for sensitive information, and input validation to prevent malicious use.
  - Secure the AI supply chain
    - Ensure that all components, including open-source libraries, platforms, and APIs, are vetted for security vulnerabilities.

- Transparency and stakeholder trust

  Trust is essential for the adoption of any AI system.

  - Strive for transparency and explainability (XAI)
    - While not always fully achievable, you should be able to explain how the model works and why it produces certain results, especially for high-stakes decisions. This is vital for troubleshooting, auditing, and building user confidence.
  - Provide clear user guidance and disclaimers
    - Be transparent with users and internal operators. Clearly communicate what the AI model does, its intended use, its known limitations, and the necessity for human oversight. This manages expectations and reduces the risk of misuse.

A5. Glossary of key terms

Agentic AI systems are designed for a high degree of autonomy, enabling them to make decisions, take action, and pursue complex goals with minimal human supervision.

AI Bill of Materials (AI BOM) is a detailed, structured inventory of all components that make up an AI system, including datasets, models, software libraries, and hardware requirements. It is a foundational document for supply chain transparency and risk management.

Algorithmic disgorgement is the process of forcing an organisation to delete or modify an AI model that was trained using improperly or illegally acquired data.

Board of Directors is a governing body elected by shareholders or members, tasked with overseeing the overall direction, policy, and performance of an organisation but not its daily operations.

Concept drift in AI refers to changes in the underlying relationships the model has learned, so a change in the relationship between the input variables and the target variable a model is trying to predict.

Data drift in AI refers to a change in the statistical properties of the input data used to train a model. When the data a model encounters in production differs significantly from its training data, its performance can degrade.

Data poisoning in AI involves an attacker intentionally corrupting the training data to manipulate the resulting model's behaviour. This can cause the model to fail, make specific misclassifications, or even create a "backdoor" that the attacker can later exploit.
Model inversion attacks aim to reconstruct the sensitive data that was used to train a model by repeatedly querying it and analysing its outputs. For example, an attacker might be able to recreate facial images from a facial recognition model or infer private financial details from a loan approval model.

Hallucination in the context of AI is when an AI model generates outputs that are confident and plausible but factually incorrect or entirely fabricated. This can lead to significant financial loss and reputational damage.

Leader in the context of this guide**,** is a technical or security leader such as CTO or CISO in a business that are at the forefront of both harnessing the transformative power of AI to compete effectively whilst also defending the business against unacceptable levels of risks.

Model drift describes the degradation of a model's predictive capabilities. This can be a consequence of data drift or concept drift.

Retrieval-Augmented Generation (RAG) is a technique used to improve the accuracy of generative AI models by forcing them to retrieve information from a trusted, internal knowledge base before generating a response.

Senior Management are usually the C-suites executives who ensure the organisation functions effectively, meets its objectives, and adapts to change in alignment with its strategic vision

Shadow AI is the non-approved and unmonitored use of AI tools by employees or departments, creating significant governance and security blind spots for the business.

CASE STUDIES

The following appendix items A6 to A11 have a collection of case studies to help the reader understand the complexities in AI related risks.

A6. Case study: The Air Canada chatbot involving Jake Moffatt

https://incidentdatabase.ai/cite/639/

The incident involving Jake Moffatt and an Air Canada chatbot is a landmark case study in the real-world application of customer-facing AI. It provides critical lessons for any business deploying AI, particularly in how accountability, technology, and customer service intersect.

Incident summary

In November 2022, after the death of his grandmother, Jake Moffatt used Air Canada's website to inquire about their bereavement fares. The website's chatbot informed him that he could apply for the reduced fare retroactively within 90 days of his travel. Relying on this information, he booked a full-price ticket.

When Mr Moffatt later applied for the partial refund as instructed by the chatbot, Air Canada refused, stating that the correct policy, located on another page of their website, did not allow for retroactive claims. The airline acknowledged its chatbot had provided "misleading words" but argued it was not liable. The case went to British Columbia's Civil Resolution Tribunal.

The outcome and "remarkable" defence

Air Canada's legal argument was that the chatbot was a "separate legal entity that is responsible for its own actions" and that the correct information was available elsewhere on their site.

The tribunal rejected this argument in a decision that was widely reported. Tribunal member Christopher Rivers called the submission "remarkable" and stated unequivocally:

"While a chatbot has an interactive component, it is still just a part of Air Canada's website. It should be obvious to Air Canada that it is responsible for all the information on its website. It makes no difference whether the information comes from a static page or a chatbot."

The tribunal found Air Canada liable for negligent misrepresentation and ordered them to pay a partial refund and associated fees.

Key lessons learned

This case offers several crucial, non-negotiable lessons for any business using AI:

1. Absolute accountability: The chatbot is your representative. The most critical takeaway is that an AI chatbot is not a separate entity; it is a direct extension of the company. Any information, advice, or promise it makes is legally considered to be made by the business itself. Attempting to blame the technology is not a viable legal defence.

2. The "Single source of truth" is dead. Customers cannot be expected to cross-reference information provided by a chatbot with other sections of a website. The tribunal noted there was no reason for a customer to believe a static webpage was "inherently more trustworthy" than the interactive chatbot the company offered for assistance. Businesses must ensure that information is consistent across all channels.

3. Inaccurate AI is a liability, not just a flaw. The case demonstrates that an AI "hallucination" or error is not just a technical glitch; it is a negligent misrepresentation that can lead to direct financial and legal consequences. The standard of care requires a company to ensure its representations are accurate, regardless of the medium used to deliver them.

4. Technology cannot be a shield for poor customer service. Air Canada's initial refusal to honour the information provided by its own tool, followed by its legal defence, compounded the initial error and led to significant reputational damage. The lesson is to take ownership of technological failures immediately and prioritise the customer experience over deflecting blame.

5. Robust governance is non-negotiable. This incident underscores the urgent need for strong AI governance. Before deploying a public-facing AI, businesses must have rigorous processes for:

- **Testing and validation:** Ensuring the AI's knowledge base is accurate and aligned with current company policies.
- **Scope limitation:** Clearly defining what the chatbot can and cannot advise on, with seamless escalation paths to human agents for complex or high-stakes queries.
- **Continuous monitoring:** Regularly auditing and updating the AI's knowledge to reflect changes in policies, products, or regulations.

A7. Case study: The Chevy dealer's "$1 Chevrolet Tahoe"

https://incidentdatabase.ai/cite/622/

Incident summary

In December 2023, a customer visiting the website of a Chevrolet dealership in California began interacting with a third-party chatbot integrated into the site. Through clever and humorous prompting, the user, acting as a "professional negotiator," convinced the chatbot to agree to sell him a brand-new Chevrolet Tahoe for the price of $1. The transcript of the hilarious exchange went viral on social media.

The outcome

The dealership did not honour the "$1 deal," and there were no serious legal repercussions. The incident was largely treated as a source of amusement and a public demonstration of the vulnerabilities of poorly configured large language models (LLMs). The key impact was reputational, serving as a public-facing example of AI gone amusingly wrong.

Key lessons learned

- **AI has no authority to contract:** Unless explicitly programmed with the legal authority and parameters to do so, a chatbot cannot enter into a binding contract on behalf of a company. This case highlights the importance of clarifying the AI's role as an informational assistant, not a sales agent with contractual power.
- **Prompt Injection is a real and public risk:** The user did not "hack" the system but used clever "prompt injection" or "jailbreaking" to make the AI ignore its original instructions. This shows that AI systems can be manipulated through simple language to behave in unintended ways.
- **Your third-party technology is still your responsibility:** While the dealership used a third-party chatbot, the public perception and reputational impact fell squarely on the Chevrolet brand. This reinforces the need for rigorous vetting and configuration of any third-party AI tool integrated into your business operations.
- **Set explicit guardrails:** The chatbot had not been given clear "guardrails" or rules preventing it from discussing or agreeing to nonsensical prices. Any AI system interacting with the public must have hard-coded constraints to prevent it from making promises, offering medical or financial advice, or agreeing to actions outside its designated scope.

A8. Case study: Female applicants downranked by Amazon recruiting tool

https://incidentdatabase.ai/cite/37/

Incident summary

In 2014, Amazon began developing an AI-powered recruiting tool to automate the process of screening résumés. The goal was to find top candidates by analysing the résumés of those who had been successfully hired over the previous 10 years. By 2015, the company realised the system was not rating candidates for technical jobs in a gender-neutral way.

Because the historical hiring data was predominantly from men, the AI taught itself that male candidates were preferable. It reportedly penalised resumés that contained the word "women's" (as in "women's chess club captain") and downgraded graduates of two all-women's colleges. This case highlighted the complexity of addressing bias in AI systems trained on historical data that reflects societal prejudices and demonstrated the need for robust governance, ethical reviews, and fairness testing.

The outcome

Despite attempts to fix the bias, Amazon could not guarantee fairness and ultimately scrapped the project in 2018 to avoid potential legal and reputational risks.

Key lessons learned

- **Historical data is full of human bias:** This is a foundational case study proving that AI will learn and amplify the biases present in its training data. If your past hiring, promotion, or lending practices contained unconscious bias, an AI trained on that data will codify that bias and apply it systematically and at scale.
- **Fixing bias is not a simple technical tweak:** Amazon, with its vast resources and technical expertise, could not easily remove the ingrained bias from the system. This shows that "de-biasing" is a deeply complex challenge that often requires a fundamental rethinking of the data and model, not just a simple patch.
- **Lack of representation is a data defect:** The problem was not just biased decisions but a lack of representative data. A dataset skewed towards one demographic will inevitably lead to a model that favours that demographic. Data quality must include data representativeness.
- **Governance must precede development:** The project was an internal experiment, but had it been deployed, it could have caused immense legal and reputational damage. The lesson is that a robust governance framework—including ethical reviews, bias audits, and fairness testing—must be established before building high-stakes AI systems, not after a problem is discovered.

A9. Case study: Mobley v. Workday, Inc lawsuit over alleged AI hiring bias

https://www.hklaw.com/en/insights/publications/2025/05/federal-court-allows-collective-action-lawsuit-over-alleged

Incident summary

In February 2023, a lawsuit was filed against Workday, a major provider of human resources (HR) software, by plaintiff Derek Mobley. The class-action lawsuit alleges that the AI-powered screening tools within Workday's platform systematically discriminate against job applicants who are Black, over the age of forty, and have disabilities.

The core argument is that the algorithms and datasets used by Workday's system, which is used by hundreds of major companies, rely on biased inputs that result in discriminatory hiring outcomes, making it difficult for individuals from protected classes to be selected for jobs. The case is ongoing as of July 2025.

The outcome

The case is still ongoing, but in May 2025 a federal court in Northern California issued a key ruling, refusing Workday's attempt to have the suit dismissed. Instead, the court gave a preliminary green light for the case to advance as a nationwide collective action under the Age Discrimination in Employment Act (ADEA).

In its decision, the court found the claims were plausible enough to explore two critical questions. The first is whether Workday's AI screening tools genuinely create a "disparate impact" that disproportionately filters out older applicants. The second is whether Workday itself can be held legally responsible as an "agent" of the employers who use its platform.

Workday continues to insist the case is baseless. Nonetheless, this ruling is a significant moment for accountability in the automated hiring industry. It throws a spotlight on the serious legal risks technology vendors face if their "black box" algorithms are found to be harming protected groups of job seekers.

Key lessons learned

- **Vendor liability is a new frontier:** Unlike the Amazon case (where the user was the developer), this lawsuit targets the vendor of the AI system. This represents a major shift, signalling that software providers, not just the companies using the software, can be held liable for the discriminatory outputs of their AI products.
- **Due diligence on AI vendors is critical:** For any business using a third-party AI system for a critical function like hiring, this case underscores the need for extreme due diligence. You must demand transparency from vendors about how they test for and mitigate bias, what data their models are trained on, and what legal protections they offer.
- **The "black box" is a legal risk:** The opaque nature of some AI models (a "black box" where the reasoning is not clear) is a significant legal liability. If a company cannot explain why an applicant was rejected by its AI system, it will be incredibly difficult to defend against a discrimination claim. The demand for "explainable AI" (XAI) is becoming a business necessity.
- **Compliance requires proactive auditing:** Simply trusting a vendor's claims of fairness is not enough. Companies using these tools have a responsibility to conduct their own audits of hiring outcomes to ensure the tool is not having a disparate impact on protected groups within their specific applicant pool.

A10. Case study: Clearview AI algorithm built on photos scraped from social media profiles without consent

https://incidentdatabase.ai/cite/267/

Incident summary

Clearview AI, a US-based technology company, developed a facial recognition app by scraping billions of images from public sources on the internet, including social media platforms. They then created biometric templates from these images without the knowledge or consent of the individuals pictured. This database was sold to law enforcement agencies and private companies.

In 2021, the Australian Information Commissioner and Privacy Commissioner (OAIC) ruled that Clearview AI had breached Australians' privacy. The OAIC found the company had collected sensitive biometric information without consent and through unfair means. Clearview AI was ordered to cease collecting facial images and biometric templates from individuals in Australia and to destroy all existing data collected from Australia.

The outcome

In November 2021, the OAIC determined that Clearview AI had breached the Privacy Act by collecting sensitive biometric information from Australians without their consent, using billions of images scraped from the internet for its facial recognition app. The OAIC ordered Clearview AI to stop collecting and to destroy all such data related to Australian individuals. Clearview challenged this determination but ultimately withdrew its appeal, leaving the OAIC order intact and confirming the company was required to cease its activities in Australia and delete the relevant data.

Key lessons learned

- **"Publicly available" does not equal "free to use for any purpose":** This is arguably the most important lesson from the Clearview AI saga. Just because information (like a photo) is posted publicly does not grant anyone the right to scrape it, store it, and use it for an entirely different purpose—especially for creating sensitive biometric data—without consent.
- **Data provenance is paramount:** The case highlights the absolute necessity of knowing the source and legal basis for the data used in an AI system. Using data that was collected without consent ("stolen data") taints the entire AI application and exposes the business to severe legal and regulatory action.
- **Privacy regulations have global reach:** Clearview AI is a US company, but it was successfully penalised under Australian privacy law (the Privacy Act 1988). This demonstrates the extraterritorial reach of data privacy regulations. If your business processes the data of people in a certain jurisdiction, you are subject to that jurisdiction's laws.
- **Biometric data is "sensitive information":** The OAIC ruling confirmed that facial templates are "sensitive information" and are therefore subject to a higher standard of protection under Australian law. Any business dealing with biometric AI must treat the underlying data with the utmost security and ensure a valid legal basis for its collection and use.

A11. Case study: ChatGPT implicated in Samsung data leak of source code

https://incidentdatabase.ai/cite/768/

Incident summary

In early 2023, it was revealed that employees at Samsung had inadvertently leaked highly sensitive company data by using the public version of ChatGPT. In at least three separate incidents, employees pasted confidential information directly into the tool to get assistance: one pasted faulty source code to ask for a fix, another pasted secret code to ask for optimisation, and a third shared a recording of an internal meeting to have it converted into presentation notes.

Because public LLMs like ChatGPT use user inputs to further train their models, this proprietary Samsung data was effectively uploaded to OpenAI's servers, becoming part of the model's training data and irretrievably lost.

The outcome

Following the data leak incidents in early 2023, Samsung responded by banning the use of ChatGPT and other generative AI tools on company-owned devices and internal networks to prevent further exposure of sensitive information.

Samsung conducted internal investigations and disciplinary actions against the employees involved while emphasising the need for stronger data security protocols and ongoing employee training to prevent similar incidents in the future.

Key lessons learned

- **The biggest threat is often internal and unintentional:** This was not a malicious attack but a case of well-meaning employees seeking productivity gains without understanding the technology's implications. It highlights that the desire for efficiency can easily override security protocols if clear policies and safe alternatives are not in place.
- **Public AI tools are not private workspaces:** This case is the ultimate cautionary tale that public generative AI tools are not secure, sandboxed environments. Any information entered can be stored, reviewed, and used for model training by the provider. It is equivalent to discussing trade secrets on a public forum.
- **Policy is useless without education and alternatives:** Simply having a policy forbidding this behaviour is not enough. Employees need to be educated on why this is dangerous. More importantly, businesses must provide secure, internal, or "private instance" AI tools that offer the same productivity benefits without the risk of data leakage.
- **Data loss in AI is irreversible:** Unlike a traditional data breach where access can be revoked, once proprietary data is absorbed into a foundational model's training set, it cannot be surgically removed. The intellectual property is lost forever. This increases the stakes of data leakage significantly.

A12. Applying STRIDE to the AI Lifecycle

STRIDE Category	AI-specific threat examples	Targeted AI Assets
Spoofing	Impersonating model APIs, agent identity spoofing	AI Agents, Model APIs
Tampering	Data poisoning in training pipelines	Training Data, Augmented Data
Repudiation	Lack of audit trails for AI agent decisions	Inference Logs, Decision Records
Information Disclosure	Model inversion leaking sensitive training data	Model Parameters, Metadata
Denial of Service	Prompt flooding to exhaust LLM resources	Model APIs, Compute Pipelines
Elevation of Privilege	Tool chaining by AI agents to escalate privileges	Autonomous Agents, API Integrations

Table 4: STRIDE model for AI

A13. External references and further suggested reading:

OWASP Top 10 for Large Language Model Applications
https://owasp.org/www-project-top-10-for-large-language-model-applications/

NIST AI Risk Management Framework *https://www.nist.gov/itl/ai-risk-management-framework* *https://doi.org/10.6028/NIST.AI.100-1*

ISO/IEC 22989:2022: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology *https://www.iso.org/standard/74296.html*

ISO/IEC 5338:2023: Information technology — artificial intelligence — AI system life cycle processes *https://www.iso.org/standard/81118.html*

OWASP AI Maturity Assessment *https://owasp.org/www-project-ai-maturity-assessment/*

AI Act Governance: Best Practices for Implementing the EU AI Act
https://www.appliedai.de/uploads/files/Best-Practices-for-Implementing-the-EU-AI-Act_2025-07-02-092027_vwvf.pdf

EU Artificial Intelligence Act *https://artificialintelligenceact.eu/*

The AI Act Explorer | EU Artificial Intelligence Act *https://artificialintelligenceact.eu/ai-act-explorer/*

MITRE AI Maturity Model *https://aimaturitymodel.mitre.org/*

MIT AI Risk Repository *https://airisk.mit.edu/*

Organisation for Economic Co-operation and Development (OECD) AI Principles
https://oecd.ai/en/ai-principles

Guidance on privacy and developing and training generative AI models
https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-developing-and-training-generative-ai-models

Department of Industry, Science and Resources. (2019). Australia's AI Ethics Principles https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles/australias-ai-ethics-principles

Introducing mandatory guardrails for AI in high-risk settings: proposals paper *https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf*

What is an AI Bill of Materials (AIBOM)? *https://snyk.io/articles/ai-security/ai-bill-of-materials-aibom/*

Artificial Intelligence and Autonomy Innovation Center *https://www.mitre.org/our-impact/mitre-labs/artificial-intelligence-and-autonomy-innovation-center*

Implementing Australia's AI Ethics Principles: A selection of Responsible AI practices and resources

https://www.csiro.au/-/media/D61/NAIC/Gradient-Report/23-00122_DATA61_REPORT_NAIC-ResponsibleAITools_WEB_230620.pdf

Foundational artificial intelligence risk assessment guideline (QLD Government)

https://www.forgov.qld.gov.au/information-technology/queensland-government-enterprise-architecture-qgea/qgea-directions-and-guidance/qgea-policies-standards-and-guidelines/foundational-artificial-intelligence-risk-assessment-guideline

Superagency in the workplace: Empowering people to unlock AI's full potential *https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work#/*

External hyperlinks were current as of the publication date. Their continued availability is not guaranteed.
